科技文献中技术关联自动发现方法研究

■ 徐珍珍　张均胜　刘文斌
中国科学技术信息研究所　北京 100038

摘　要：[目的/意义]技术关联分析不仅有助于科技管理部门做好规划布局、企业机构在技术研发方面强优补短，还能帮助科研人员选择技术创新方向和路径。[方法/过程]提出一种基于科技文献文本分析构建问题－技术矩阵并用于发现技术关联的方法。首先，提取文献中研究问题及对应技术，形成问题－技术矩阵；然后，利用技术相似度进行技术项合并聚类以应对数据稀疏性问题，进而分析建立技术之间的关联；最后，进行实验以验证该方法的可行性。[结果/结论]从文献中的研究问题－方法矩阵出发对技术间关系进行归纳和总结，提出一种自动化建立技术之间的关联关系的方法，可以有效辅助发现技术关联关系，如直接竞争关系、合作关系、间接竞争关系和合作关系等。本研究是面向科技文献的技术关联发现与应用的探索性研究，研究结果可为科技创新研究与管理提供参考。

关键词：问题－技术矩阵　技术关联　技术相似度
分类号：G255.51
DOI：10.13266/j.issn.0252－3116.2021.20.012

1 引言

科技创新可以促进新产品、新服务和新商业模式的产生[1]，是经济长期增长的重要源泉之一。当前面临新的科技变革和产业变革，全球科技创新竞争日趋激烈。随着中国科技水平的不断提高，我国与世界发达国家的差距在许多领域正在逐步缩小，部分领域甚至已经处于领先地位。《中国战略性新兴产业发展报告》中提到我国战略性新兴产业的创新必须要向基础性创新、引领性创新转型，要加强前瞻性基础研究、应用性基础研究，突出关键共性技术、前沿引领技术、现代工程技术和颠覆性技术创新[2]。分析技术之间的关联关系有助于了解相关技术的发展脉络，辅助选择技术创新方向及路径，对于政府科技管理、企业机构技术研发以及个人科技创新选题都有重要意义。

科技文献是分析技术关联的重要信息资源与依据。基于科技文献内容分析建立研究问题和技术之间的关联关系是科技情报分析的一项重要任务。科研人员需要不断阅读相关领域科技文献，才能了解所在领域研究工作前沿进展。海量科技文献记录了科技发展历史中绝大多数的研究问题、技术方法和实验结果。科研工作者需要及时了解科技文献作者用什么技术解决什么问题得到什么结论，建立研究问题及技术发展脉络，形成研究领域态势理解。科技文献摘要蕴含主要的研究结论，由于篇幅所限缺少研究过程详细信息，而科技文献正文包含具体过程，相关研究和实验部分包含解决研究问题的相关技术及结果比较分析，可以用于挖掘技术之间的关联。面对已存在的海量科技文献以及持续新发表的科技文献，传统的人工全文阅读文献方式日益难以应对科技文献信息过载问题，迫切需要借助人工智能等先进信息技术快速获取并分析了解科技文献内容。

本文从科技文献中抽取研究问题和技术词语构建问题－技术矩阵，进而分析发现技术之间的关联关系，为技术发展趋势预测奠定基础。研究成果将有助于科研人员开展技术创新及应用研究，还有助于科研管理者了解技术研究现状、发展脉络和最新态势，科学制定发展战略规划并正确选择技术路径。

2 相关研究

科技文献调研能够帮助科研人员了解领域发展状

＊ 本文系国家重点研发计划项目“颠覆性技术识别理论、方法与专家预判系统”（项目编号：2019YFA0707201）和中国科学技术信息研究所创新研究基金“互联网虚假科技信息识别方法研究”（项目编号：MS2021－05）研究成果之一。
作者简介：徐珍珍（ORCID：0000－0002－9641－8334），硕士研究生；张均胜（ORCID：0000－0001－8740－2851），研究员，博士，通讯作者，E－mail：zhangjs@istic.ac.cn；刘文斌（ORCID：0000－0002－5411－544X），硕士研究生。
收稿日期：2021－04－14　修回日期：2021－07－19　本文起止页码：113－122　本文责任编辑：杜杏叶

况以及发展趋势。科技文献研究正在从以元数据研究为主向深入科技文献内容分析及应用研究发展。吴江宁等[3]基于管理科学领域核心论文分析论文中的题目、摘要、关键词和发表时间等信息，利用 LDA 模型提出了科技论文的时序主题链构建方法。王佳琪等[4]通过研究科技文献元数据建立科研事件的语义链接以形成科研事件网络支撑科技情报分析。张昱等[5]对科技文献中的元数据和文本内容进行分析，构建了科技文献的语义关联网络，支持科技信息的语义浏览与检索。王艳艳等[6]对科技文献中的摘要进行分析，构建问题-方法矩阵来探索科技文献研究内容新颖性评估方法，辅助科技查新应用。J. Li 等[7]从摘要中进行信息抽取并利用 Author-Conference-Topic 模型构建学术社交网络。

技术功效矩阵默认为专利功效矩阵[8]，通过专利文献反映的主题技术方案和主要技术功能之间的特征研究来揭示技术和功效二者的关系[9]，将技术领域的技术手段与对应实现的技术功效种类构成矩阵[10]。技术词是专利组件名称、技术流程、技术方法名称、涉及的设备材料名称等[11]，通常出现在专利名称或摘要字段中；功效词是专利实施后所能表示的性能、用途、目标等，通常出现在专利摘要、发明改进等字段中[12]。翟东升等[13]提出了基于 SAO 结构和词向量的专利技术功效图构建方法，为技术功效图的自动化构建提供新的思路。专利功效矩阵可以用于分析技术主题关联、挖掘核心专利[14]、分析技术机会[15]等。

对于技术关联性目前没有一个标准的定义，不同学者根据其研究背景对技术关联性有不同的理解。栾春娟等[16]从可计量的意义上将其定义为在一个特定技术体系或技术领域内，一种技术类型与其他技术类型相关联的数量和程度。丰雷等[17]从产业集群方面认为技术关联性指技术之间的相互交叉、联动影响的关系。姜红[18]从技术系统的角度认为技术关联性是指转移技术与当地其他产业技术的交叉影响关系。P. Hofmann 等[19]认为文本技术特定语料库之间的相似性可以表明技术之间的相关性。

对技术关联分类标准主要有从技术关联对企业或产业发展的影响、技术间关联程度的测度以及从研究对象如专利出发进行分类。冯秀珍等[20]从技术用途角度将技术分为直接关联、间接关联和关键关联，之后通过分析技术关联定义了五种技术群形态以此来把握当前技术趋势，判断技术发展前景。丰雷[21]从产业集群方面认为企业间的技术关联包括纵向关联和横向关联。陈卫静等[22]以某技术领域的关键词共现网络为研究载体，通过分析某一技术在整个网络中的位置和角色，发现技术间的依存关系、互补关系、控制关系。黄斌首先从专利角度分析技术间的关联形式为直接联系和间接联系[23]，之后又进一步围绕技术关联的影响和结构及其关系测度三个方面[24]对技术之间的关联进行研究，在技术关联影响分析方面，将技术关联影响类型划分为直接关联影响和间接关联影响以及无关联影响类型；在技术关联测度方面区别了技术之间的对称和不对称关联。杨冬敏[25]从四个维度将技术关联分为 12 类，是目前已知分类较为详尽的分类框架，但是对于技术关联分类发现只停留在人工阶段。目前大多数学者从专利、企业角度对技术关联进行研究，一项技术经常与其他技术结合实现特定功能。

通常技术条目表现为单词或者短语，为减少技术分析的数目可以将技术项合并。王燕鹏[26]利用动态主题模型方法构建完整的科研机构研究主题分布及演化流程。李勇敢[27]改进了杰卡德相似系数以及杰卡德距离在共现分析中的不足，提出了相对技术相似度和相对技术关联度的概念。黄晓斌等[28]利用向量余弦相似度和平方欧式距离两种相似度的标准对通讯技术领域进行聚类分析。包翔等[29]尝试将概率潜在语义分析运用到专利文本的标引中，为技术主题聚类提供了一种新的思路。汪锦霞等[30]对主题词进行聚类展现技术发展的脉络细节。

技术关联关系已有相关研究对技术关联的分类大多较为笼统，如强关联、弱关联，少有研究关注技术间的具体关联形式。目前已知分类最详尽的技术关联分类框架采用的是人工标注方式耗时长且主观性强，因此本文尝试从研究问题角度切入自动发现科技文献中技术之间的关联关系。

3 构建问题-技术矩阵

3.1 相关概念

本文中，研究问题(question)指科技文献针对的研究问题。技术(technology)指为了解决研究问题所提出的技术。问题-技术矩阵(question-technology matrix, QTM)指以研究问题和技术为两个维度构建的矩阵，用于描述研究问题和技术之间的对应关系。技术关联指以研究问题为纽带而形成的技术之间的关系。杨冬敏将技术关联分为 12 类，从技术应用角度分有竞争关系和互补关系，从技术体系结构分有包含关系。本研究尝试自动化发现上述技术关联，依据研究问题与技

术之间的对应关系发现技术之间的关联,具体分为直接竞争关系、间接竞争关系和合作关系和包含关系。

直接竞争关系是指如果两个或两个以上的技术都解决同一个问题,则两个或两个以上技术之间为直接竞争关系。图1(1)中技术1和技术2为直接竞争关系。

合作关系是指如果两个或两个以上的技术都解决不同的问题,但是它们之间通过组合共同解决某一个问题,则两个或两个以上技术之间存在合作关系。图1(2)中技术3和技术4为合作关系。

间接竞争关系是通过合作关系体现的,如果两个或两个以上技术通过合作可以解决某个问题,同时又有其他技术可以解决该问题,那么该技术与这两个或两个以上技术之间形成间接竞争关系。图1(3)中技术3和技术8、技术4和技术8也为间接竞争关系。

包含关系是指如果两个或两个以上的技术合并成一项技术或者是聚成一个技术类,那么合并后的技术或者技术类与这两个或两个以上技术形成包含关系。包含关系的推理规则如下,$T_1 \xrightarrow{a} T_2, T_2 \xrightarrow{a} T_3 \Rightarrow T_1 \xrightarrow{a} T_3$,其中$T_1$、$T_2$、$T_3$可以是单独技术点或技术群,a表示技术之间的包含关系。图1(4)中技术类T_1包含技术5和技术6,技术类T_2包含技术类T_1、技术7,则技术类T_2也包含技术5和技术6。

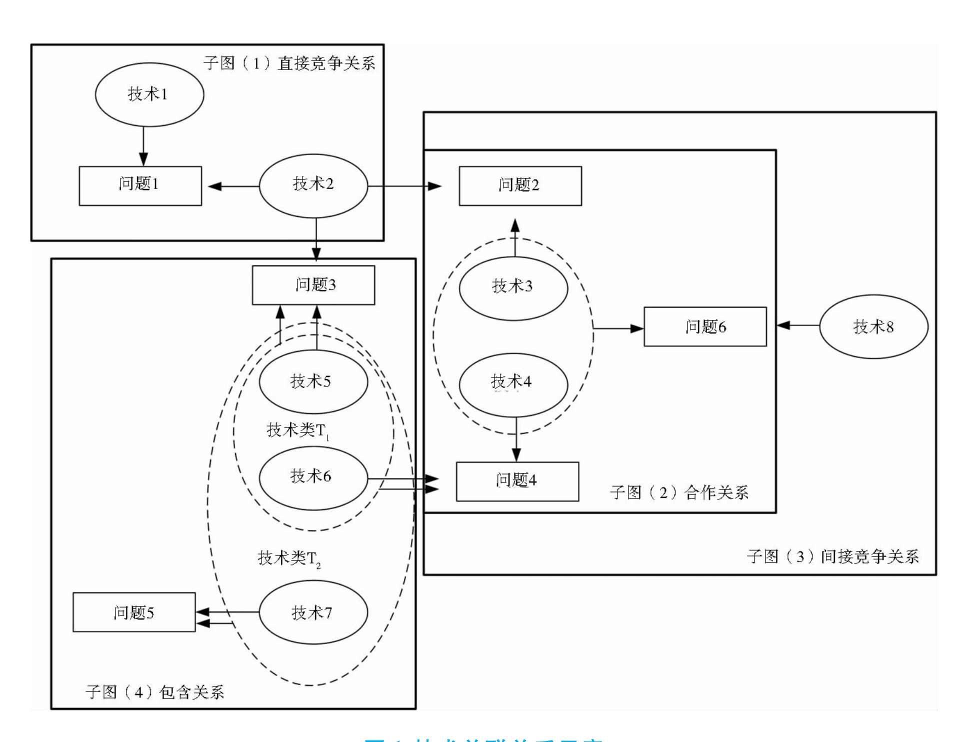


图1 技术关联关系示意

3.2 问题－技术矩阵构建及降维方法

构建问题－技术矩阵是技术关联发现的基础。不同技术之间可能存在多种关联,以研究问题为纽带可以将技术及其关联形成技术关联网络。科技文献中,研究问题和技术通常出现在科技文献的摘要(abstract)、介绍(introduction)、相关研究(related works)和实验(experiments)等章节部分。

图2是基于科技文献构建问题－技术矩阵的流程图,具体步骤如下:

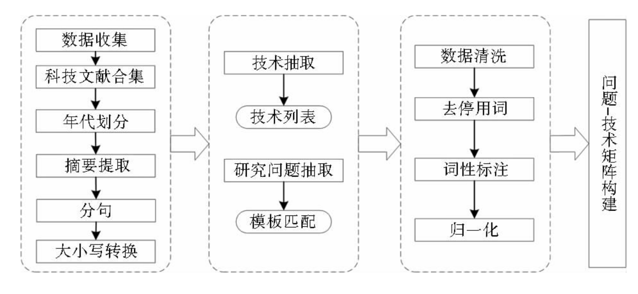


图2 矩阵构建流程

科技文献集合$D = \{d_1, d_2, \cdots\cdots, d_k\}$,其中$k$为科技文献的数量,$k>0$。对科技文献文本进行分句、大小写转换处理,构成科技文献句子集合$S = \{s_1, s_2, \cdots\cdots, s_n\}$,$n$为分句的数量,$n>0$。

从每句中s_n抽取问题与技术,构成技术集合$T = \{t_1, t_2, \cdots\cdots, t_x\}$,研究问题集合$Q = \{q_1, q_2, \cdots\cdots, q_y\}$,其中$x$为技术数量,$y$为研究问题数量,$x>0$,$y>0$。

构建 $x*y$ 阶问题 - 技术矩阵 QTM，$QTM(x,y)=1$ 表示技术 t_x 可以解决问题 q_y；$QTM(x,y)=0$ 则表示技术 t_x 不能解决问题 q_y。

技术词和研究问题的确定以及矩阵单元项的填充是问题 - 技术矩阵构建的难点。已有研究大多用关键词或主题词来代表文献所用技术，主题词表达技术点的粒度较大。本研究通过技术列表匹配的方式抽取技术词，而技术列表由领域专家整理而成，与关键词相比能够更准确的描述技术，与主题相比表达的粒度更小。科技文献中摘要是结构化的，表达形式较为规范且具有一定的规则，因此本研究采用基于模板匹配的方式来抽取摘要中的问题。在问题 - 技术矩阵构建时需对科技文献的摘要进行扫描分析，当技术词语与问题词语在同一句时，本文假定该技术可以解决这个问题，在矩阵单元项填充为 1，否则为 0。

以英文科技文献为例，对研究问题的描述模式梳理如表 1 所示，根据研究问题在句子中出现位置不同，将研究问题的模板分为两大类，共 23 个模板。由于模板归纳总结受语料库限制，模板库需要结合语料库内容进行更新维护。

表 1 问题模板及样例

类型	模板	样例(AAAI 数据集中统计所得)
类型 1	[Tt]o (.*?) issue	To alleviate the sparsity issue
	(.*?) (is|are) (.*?) issues? \b	the level of granularity is still a critical issue
	deal with (.*?) problems? \b	Thus these methods cannot deal with the data sparsity without commonly rated items (DS-WO-CRI) problem
	tackle the problems? \b (.*?) $	we also tackle the problems arising from noise and variation in microblogging texts
	(.*?) (is|are) important (to|for) (.*?) $	Image localization is important for marketing and recommendation of local business
	(.*?) (has|have|had) become (.*?) topics? \b	How to efficiently share the underlying information and knowledge
	(.*?) (has|have|had) drawn (.*?) attention	Detection of overlapping communities has drawn much attention
	(.*?) (has|have|had) been investigated	Top-N recommender systems have been investigated widely both in industry and academia.
	(.*?) (has|have|had) attracted (.*?) research	Sentiment classification on Twitter has attracted increasing research in recent years
	(.*?) becomes? \b (.*?) tasks? \b	location-based social networks (LBSNs) becomes a significant task
	(.*?) plays? \b (.*?) role	what causes traffic accident and early alarms for some possible ones will play a critical role on planning effective traffic management
	(.*?) becomes? \b (.*?) problems? \b	the number of traffic accidents have significantly increased globally over the past decades and become a big problem for human society
	solve (.*?) problems? \b	We hypothesize that most poker games can be solved as a pattern matching problem
	the problems? \b of (.*?) $	we address the problem of personalized next Point-of-interest (POI) recommendation
	study (.*?) problems? \b	We study the Maximum Weighted Matching problem in a partial information setting
	(.*?) (is|are) (.*?) task	The design of the best economic mechanism for Sponsored Search Auctions (SSAs) is a central task in computational mechanism design/game theory
	considers? \b (.*?) problems? \b	We consider the following problem in which a given number of items has to be chosen from a predefined set
类型 2	propose (.*?) to (.*?) $	we propose a deep convolutional neural network architecture, MUST-CNN to predict protein properties.
	(.*?) task for (.*?) $	we study a challenging task for integrating users' information from multiple heterogeneous social networks
	(.*?) impact (.*?) $	Their presence in user homepage stream of news aggregator sites may adversely impact user experience
	solve (.*?) problems? \b (.*?) $	But existing online hashing methods still cannot solve two essential problems: efficient updating of hash codes and analysis of cross-modal correlation
	[Aa] (.*?) question (.*?) $	A natural and well-studied question is the tournament fixing problem (TFP): given the set of all pairwise match outcomes
	problem for (.*?) $	it is a much more complex problem for imperfect information games

初始构建的问题 - 技术矩阵特征维度高，特征向量稀疏，不利于技术关联的发现与分析，因此本文通过合并技术项目来降低矩阵的稀疏性。通过合并技术项和问题项，形成新的 $n*m$ 阶问题 - 技术矩阵 QTM'，$(n>0, m>0)$。合并遵循如下规则，假设技术 T_x、T_y 可以合并，并且合并后的技术为 T'_z：

①如果 $QTM(i_x,j)=0$，$QTM(i_y,j)=0$，则合并后的 $QTM'(i_z,n)=0$；

②如果 $QTM(i_x,j)=0$，$QTM(i_y,j)=1$，则合并后的 $QTM'(i_z,n)=1$；

③如果 $QTM(i_x,j)=1$，$QTM(i_y,j)=0$，则合并后的 $QTM'(i_z,n)=1$；

④如果 $QTM(i_x,j)=1$，$QTM(i_y,j)=1$，则合并后的 $QTM'(i_z,n)=1$；

同样地遵循如下问题合并原则，假设问题 Q_x、Q_y 可以合并，并且合并后的问题为 Q'_z：

①如果 $QTM(i,j_x)=0$，$QTM(i,j_y)=0$，则合并后的 $QTM'(m,j_z)=0$；

②如果 $QTM(i,j_x)=0$，$QTM(i,j_y)=1$，则合并后的 $QTM'(m,j_z)=1$；

③如果 $QTM(i,j_x)=1$，$QTM(i,j_y)=0$，则合并后的 $QTM'(m,j_z)=1$；

④如果 $QTM(i,j_x)=1$，$QTM(i,j_y)=1$，则合并后的 $QTM'(m,j_z)=1$；

降低矩阵稀疏性具体可分为两步：首先，科技文献中明确提到含有包含意义的词（如 include、contain、consist of 等），将这些技术条目进行合并。其次，构建技术相似度矩阵，通过层次聚类，进行技术条目合并，进一步发现技术之间的关联。

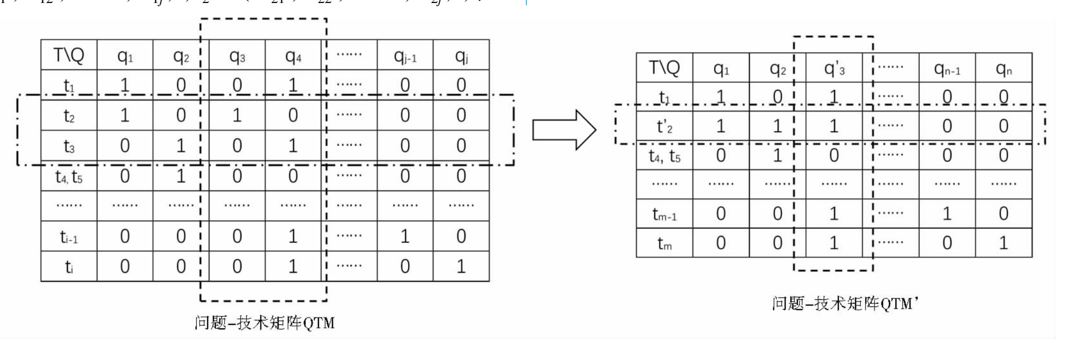
利用 word2vec 将技术词与或短语转换成词向量，再用向量间的夹角余弦值计算两两技术间的相似度，构建技术相似度矩阵。假设 word2vec 转换的技术向量为 $t_1=(w_{11},w_{12},\cdots\cdots,w_{1j})$，$t_2=(w_{21},w_{22},\cdots\cdots,w_{2j})$，在向量空间模型中，$t_1$，$t_2$ 的夹角余弦值即为技术 t_1，t_2 的相似度，相似度阈值在 0－1 之间，值越大则两个技术的相似性越大，如公式（1）所示。

$$vector_{similarity}(T_1,T_2)=cos(\theta)=\frac{\Sigma_{i=1}^{j} w_{1i} * w_{2i}}{\sqrt{\Sigma_{i=1}^{j}(w_{1i})^2} * \sqrt{\Sigma_{i=1}^{j}(w_{2i})^2}} \quad \text{公式(1)}$$

4 基于问题－方法矩阵发现技术关联

科技文献调研主要需求场景包括：①明确知道待解决的问题，需要了解到解决该问题的技术有哪些；②想对某个技术进行深入了解，需要知道该技术可以解决哪些问题，以及该技术是如何演化的。技术关联的发现有助于发现科研发展脉络，如通过分析技术关联网络的节点度数可以发现当前研究问题的研究热度，可以了解当前研究的前沿技术。

图 3 是问题－技术矩阵聚类前后对比图，矩阵 QTM 中虚线表示可以合并的技术或问题项目，矩阵 QTM' 中的虚线表示合并后的技术或问题项目。图 4 是聚类前后技术关联网络图，其中圆圈节点表示技术点或技术群，矩形表示研究问题，椭圆表示合并后的技术或问题项目，若矩阵中 $QTM(x,y)=1$，则在技术关联图中技术 t_x 指向问题 q_y。



T\Q	q1	q2	q3	q4	......	qj-1	qj
t1	1	0	0	1	......	0	0
t2	1	0	1	0	......	0	0
t3	0	1	0	1	......	0	0
t4, t5	0	1	0	0	......	0	0
......	......	......	......	......	......	......	......
ti-1	0	0	0	1	......	1	0
ti	0	0	0	1	......	0	1

问题–技术矩阵QTM

T\Q	q1	q2	q'3	......	qn-1	qn
t1	1	0	1	......	0	0
t'2	1	1	1	......	0	0
t4, t5	0	1	0	......	0	0
......	......	......	......	......	......	......
tm-1	0	0	1	......	1	0
tm	0	0	1	......	0	1

问题–技术矩阵QTM'

图 3　问题－技术矩阵聚类前后对比

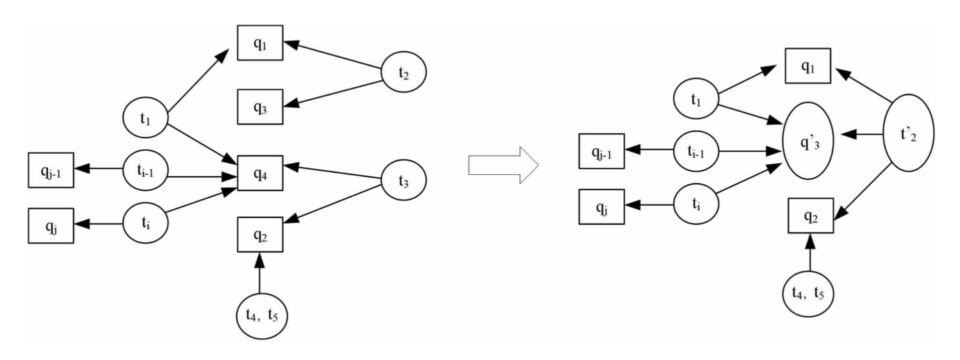


图 4　聚类前后技术关联对比

竞争关系是解决同一个问题而提出的不同技术,竞争关系的发现有利于科研人员对研究问题所用的技术进行梳理,在原有技术上进行选择和创新来解决某个问题;另外当两个或者多个技术之间存在竞争关系时,可以通过某些手段将这些技术进行融合,可能解决更多问题,从而进行技术创新。由图 3 的问题-技术矩阵 QTM,根据算法自动化发现:技术 t_1 和 t_2 可以同时解决问题 q_1,认为技术 t_1 和 t_2 是直接竞争关系,同理技术 t_1、t_3、t_{i-1} 和 t_i 可以同时解决问题 q_4,则技术 t_1、t_3、t_{i-1} 和 t_i 两两之间是直接竞争关系。合作关系的发现使原本孤立的技术之间产生了关联,为解决新问题提供了新的思考方向,科研人员可以将不同技术之间进行合理的组合,从而产生新的功效解决新的问题。由问题-技术矩阵 QTM,根据算法自动化发现,技术 t_4 和 t_5 一起发挥作用共同解决问题 q_2,那么技术 t_4 和 t_5 之间是合作关系,技术 t_4 与技术 t_3 构成间接竞争关系,同时技术 t_5 与技术 t_3 也形成间接竞争关系。包含关系的发现使得技术具有被替代潜在可能性,辅助科研人员进行颠覆性技术的发现与研究。由问题-技术矩阵 QTM 可看出,技术 t_2 可以解决问题 q_1 和 q_3,技术 t_3 可以解决问题 q_2 和 q_4,原本是两个孤立的技术,通过技术项目合并(由问题-技术矩阵 QTM' 可看出),将技术 t_2 和 t_3 合并成技术 t'_2,那么技术 t'_2 包含技术 t_2,技术 t'_2 包含技术 t_3。以下是直接竞争关系发现的算法描述:

算法名称:directcompetition

输入: 问题技术矩阵 Question-Technology Matrix

输出: 直接竞争技术列表 directcompetition_list

过程:

```
Question list row0_quetions
Col_value context
For i1, v1 in row0_quetions do :
  For i2, v2 in context[1:] do :
    If v2 = 1 then t←cell_value, t append in dt_list #矩阵单元格内容为 1 的行列值
  End for
  If length(dt_list) > 1 then
    For d1, dv1 in dt_list do :
    For d2, dv2 in dt_list do :
      If d1 < d2 then dv1, dv2 append in d_value
        If d_value not in directcompetition_list then d_value append in directcompetition_list
    End for
  End for
End for
Return directcompetition_list
```

5 实验

5.1 数据集

有研究表明,专利文献滞后于期刊论文一年,期刊论文滞后于会议论文一年[31],会议论文具有新颖性、及时性等特点,因此本文选取会议论文作为实验数据以便及时发现技术发展动向。AAAI 会议汇集了全球最顶尖的人工智能领域专家学者,研究成果是人工智能界的研究风向标来源之一,数据具有权威性且容易获取。本文采用 2016-2020 年 AAAI 会议发表的论文为数据集,在 Web of Science 核心合集中以 AAAI 为会议名称,时间范围限定在 2016-2020 年进行检索,共检索出 4 772 篇文献(检索时间为 2021 年 3 月 15 日)。

5.2 问题-技术矩阵构建

对获取的文献首先要进行预处理,之后分两步分析技术之间的关联。①提取出文本中问题和技术,研究问题的抽取式根据模板匹配得到的,技术的抽取依赖技术列表(技术列表来源于 https://www.ctolib.com/jiqizhixin-ai-terminology-page.html),对提取的问题技术进行数据处理,将缩写与全称规范化为"[全称]"形式,如将[Long Short-Term Memory]和[LSTM]规范化为[Long Short-Term Memory],用于构建问题-技术矩阵 QTM;②利用技术相似度矩阵和层次聚类构建合并后的问题-技术矩阵 QTM'。

对 4 772 篇文献中的摘要字段进行分句,共得到 12 369 句。表 2 是根据总结出的 23 个问题模板在数据集中抽出的问题数量,不同模板抽取出的问题数目相差较大,其中模板 propose (.*?) to (.*?) $、(.*?) (is|are) (.*?) task 和 the problems? \b of (.*?) $ 抽取的出的问题数目最多。对抽取出的问题、技术进行整理,技术共 459 个,抽取出的问题共 2 493 个,最终得到 459 * 2439 的问题-技术矩阵,矩阵可视化结果见图 5。其中白色节点代表技术,灰色节点代表研究问题。可以看出,初始的问题-技术矩阵中数据非常稀疏。某些类似技术只解决某个问题,很少出现多个技术解决不同问题,矩阵可视化结果呈现围绕某个问题而抱团的态势,不同技术直接联系较少,如 learn fashion attribute、landmark detection、cross-domain fashion retrieval、body shape and size based fashion suggestion 等技术只能解决问题 Visual fashion analysis 而不能解决其他问题,分析产生的原因是对于某篇科技文献,该文献的作者只提到作者想解决问题的技术而很少提及该技术能否解决其他问题。节点的大小表

徐珍珍，张均胜，刘文斌. 科技文献中技术关联自动发现方法研究[J]. 图书情报工作,2021,65(20):113－122.

示该问题或技术是当前研究热点，如 T_{34} 为技术 adaptive resonance theory 可以解决 the local similarity relationships derived from the affinity matrix 等问题。

表 2　问题模板抽出问题的数量

序号	模板	数量
1	[Tt]o (. *?) issue	179
2	(. *?) (is|are) (. *?) issues? \b	41
3	deal with (. *?) problems? \b	19
4	tackle the problems? \b (. *?) $	15
5	(. *?) (is|are) important (to|for) (. *?) $	30
6	(. *?) (has|have|had) become (. *?) topics? \b	6
7	(. *?) (has|have|had) drawn (. *?) attention	13
8	(. *?) (has|have|had) been investigated	7
9	(. *?) (has|have|had) attracted (. *?) research	12
10	(. *?) becomes? \b (. *?) tasks? \b	16
11	(. *?) plays? \b (. *?) role	101
12	(. *?) becomes? \b (. *?) problems? \b	16
13	solve (. *?) problems? \b	156
14	the problems? \b of (. *?) $	336
15	study (. *?) problems? \b	122
16	(. *?) (is|are) (. *?) task	378
17	considers? \b (. *?) problems? \b	100
18	propose (. *?) to (. *?) $	781
19	(. *?) task for (. *?) $	23
20	(. *?) impact (. *?) $	150
21	solve (. *?) problems? \b (. *?) $	92
22	[Aa] (. *?) question (. *?) $	58
23	problem for (. *?) $	54

由于抽取出的问题和技术数量较多，构建出的矩阵较为庞大，后续技术之间关联发现较为复杂，因此本文采用 2017 年数据进行研究，得到 294 * 406 的问题技术矩阵并进一步合并技术项目。利用 word2vec 将技术转换成词向量之后，计算向量之间的余弦值得到对称的技术相似度矩阵，如图 6 所示，层次聚类过程中，经过多次筛选阈值和人工检验，最终选定理想阈值为 0.71，共分为 65 类。

5.3　技术关联发现

根据实验结果，直接竞争关系有 36 585 对，合作关系有 1 586 对，包含关系有 347 对。部分技术关系见表 3，从表 3 中可以看出一个技术可以解决多个问题，如 Latent Dirichlet Allocation 可以解决问题 The Sparsity 和 Mitigating The Vanishing Gradients Problem，多个技术也可共同发挥作用解决一个问题，如 Deep Learning、Gated Attention-Based Recurrent Networks 和 Word Embedding 可以解决 Multi-Task Learning (Mtl)。直接竞争关系中，Latent Dirichlet Allocation 与 Reinforcement Learning 与 Independent Component Analysis Latent 等技术形成直接竞争关系；在间接竞争关系中，Adaptive Resonance Theory 和 Deep Learning、Gated Attention-Based Recurrent Networks 等技术形成间接竞争关系；在合作关系中，Deep Learning 和 Gated Attention-Based Recurrent Networks、Word Embedding 等形成合作关系；在包含关系中，Classifier 包含 K-Nearest Neighbours Algorithm，类 C1 包含 Quadratic Programming 和 Dynamic Programming 等技术，类别 C5 包含 Performance Measure 和 Parameter Estimation 等技术。

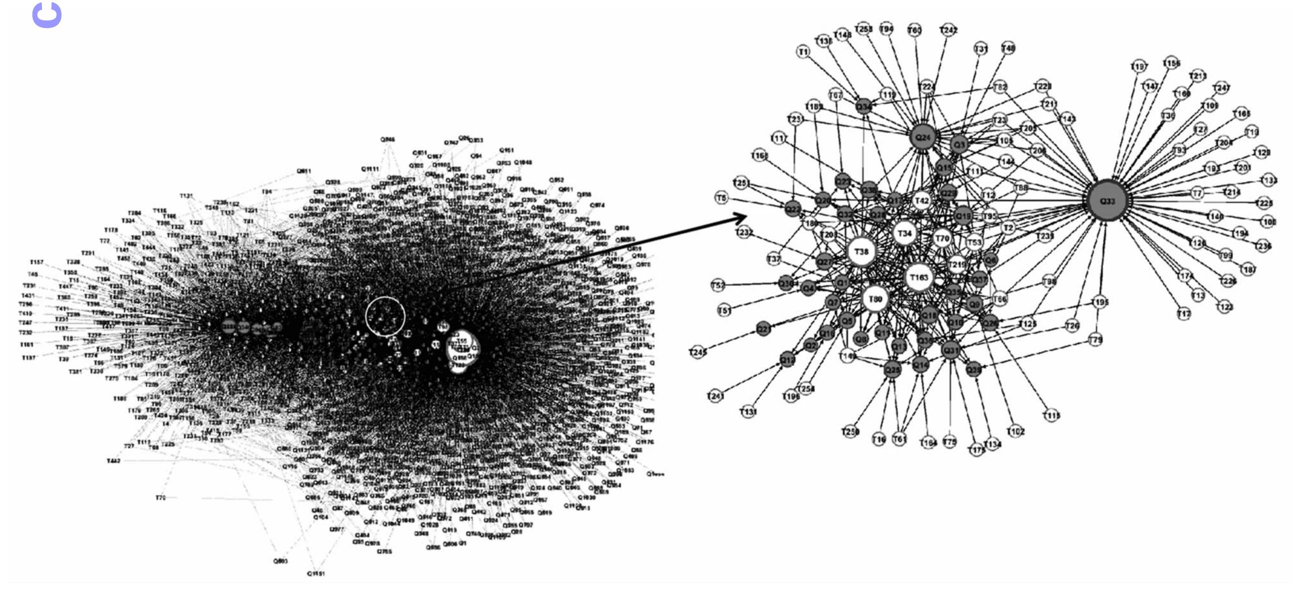


图 5　问题－技术关联

注：白色圆圈为研究技术，灰色为研究问题

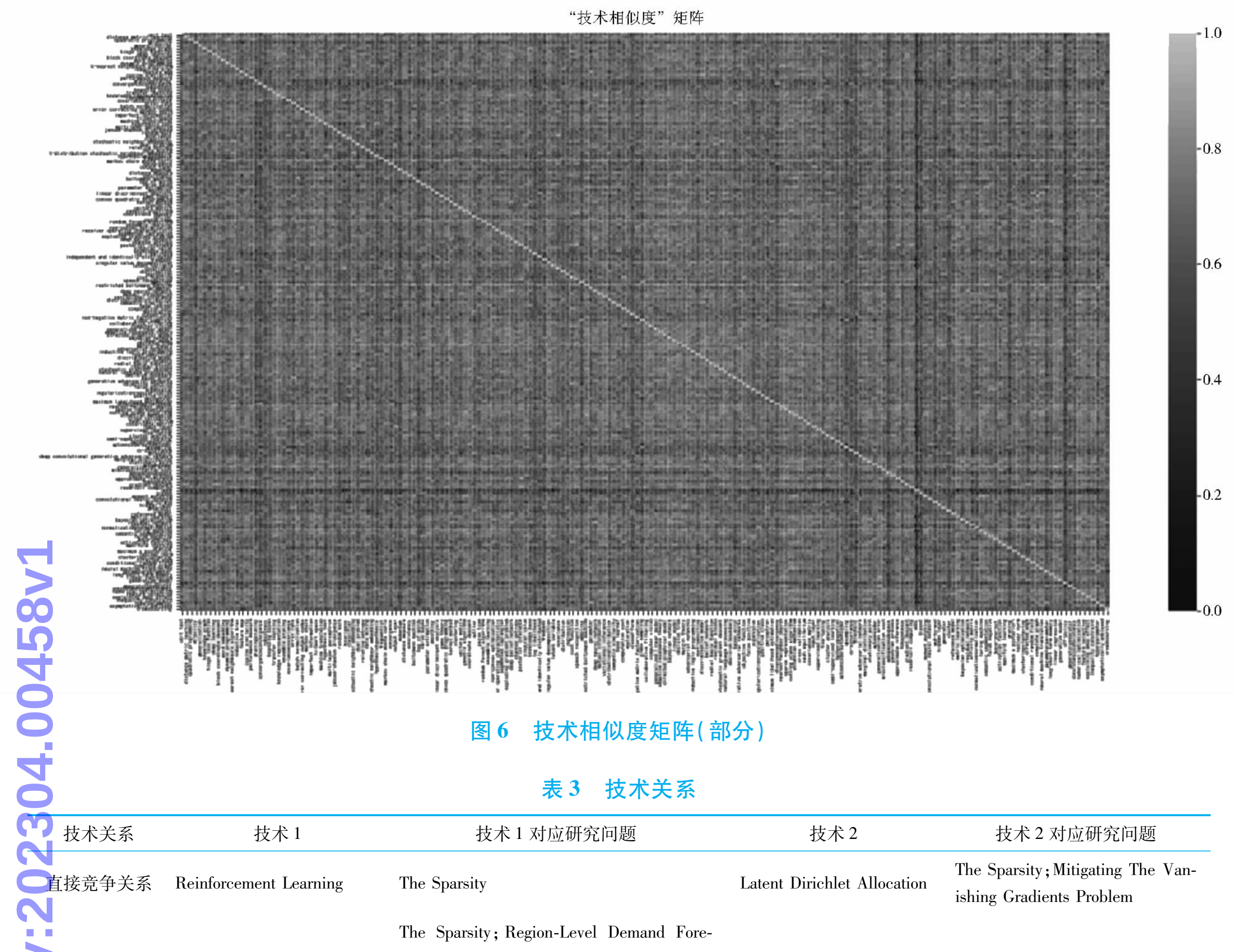


图 6 技术相似度矩阵(部分)

表 3 技术关系

技术关系	技术 1	技术 1 对应研究问题	技术 2	技术 2 对应研究问题
直接竞争关系	Reinforcement Learning	The Sparsity	Latent Dirichlet Allocation	The Sparsity; Mitigating The Vanishing Gradients Problem
	Independent Component Analysis	The Sparsity; Region-Level Demand Forecasting; Learn Context Representation To Improve Text Classification; Alleviate The Impact Of Noisy Data	Latent Dirichlet Allocation	The Sparsity; Mitigating The Vanishing Gradients Problem
	Independent Component Analysis	The Sparsity; Region-Level Demand Forecasting; Learn Context Representation To Improve Text Classification; Alleviate The Impact Of Noisy Data	Extreme Learning Machine	Alleviate The Impact Of Noisy Data
	Latent Dirichlet Allocation	The Sparsity; Mitigating The Vanishing Gradients Problem	Recurrent Neural Networks	Mitigating The Vanishing Gradients Problem
	Long-Short Term Memory	Predict Morphological Boundaries	Receiver Operating Characteristic	Predict Morphological Boundaries
间接竞争关系	Adaptive Resonance Theory	Multi-Task Learning (Mtl)	Deep Learning	Multi-Task Learning (Mtl)
	Adaptive Resonance Theory	Multi-Task Learning (Mtl)	Word Embedding	Multi-Task Learning (Mtl)
	Adaptive Resonance Theory	Multi-Task Learning (Mtl)	Gated Attention-Based Recurrent Networks	Multi-Task Learning (Mtl)
合作关系	Deep Learning	Multi-Task Learning (Mtl)	Gated Attention-Based Recurrent Networks	Multi-Task Learning (Mtl)
	Word Embedding	Multi-Task Learning (Mtl)	Deep Learning	Multi-Task Learning (Mtl)
	Gated Attention-Based Recurrent Networks	Multi-Task Learning (Mtl)	Word Embedding	Multi-Task Learning (Mtl)
	Gated Convolutional Networks	A Related Architecture	Memory Networks	A Related Architecture
包含关系	Classifier	Symmetric Positive Defined (Spd) Matrix	K-Nearest Neighbours Algorithm	Learn Depth Features Combining Local Weakly Supervised Training From Patches Followed By Global Fine Tuning With Images

（续表 3）

技术关系	技术 1	技术 1 对应研究问题	技术 2	技术 2 对应研究问题
	类别 C1	Approximate The Pareto Frontier; Integrate Semantic Attributes With Trajectories For Cross-View People Tracking; Convex Quadratic Programming	Dynamic Programming	Approximate The Pareto Frontier; Integrate Semantic Attributes With Trajectories For Cross-View People Tracking
	类别 C1	Approximate The Pareto Frontier; Integrate Semantic Attributes With Trajectories For Cross-View People Tracking; Convex Quadratic Programming	Convex Quadratic Programming	Convex Quadratic Programming
	类别 C5	The Class Imbalance ; Solve Traffic Flow Forecasting Problem; Drastically Speed Up Hlta Using A Technique Inspired By The Advances In The Method Of Moments	Performance Measure	The Class Imbalance; Solve Traffic Flow Forecasting Problem
	类别 C5	The Class Imbalance; Solve Traffic Flow Forecasting Problem ; Drastically Speed Up Hlta Using A Technique Inspired By The Advances In The Method Of Moments	Parameter Estimation	Drastically Speed Up Hlta Using A Technique Inspired By The Advances In The Method Of Moments

构建问题－技术矩阵的方法能够发现技术之间的关联,从实证分析结果来看,直接竞争关系 36 585 对,合作关系有 1 586 对,包含关系有 347 对,从结果来看直接竞争关系最多,合作关系最少,其次是包含关系,结果分布不平衡,可能的原因是:从问题出现开始,不同学者为了解决该问题提出了不同技术,直接竞争关系最多是必然,而合作关系和包含关系少是因为本文研究对象是科技文献中摘要部分,文章作者对于摘要主要提出自己的方法技术,少有提及该方法技术的发展进程以及其他技术状况。同时问题－技术矩阵的稀疏性本文采用的是层次聚类进行降维,在一定程度上满足本文需求,然而理想状态下降低矩阵稀疏性之后在很大程度上需要保留原始技术保证细粒度技术,这个方面需要在后续研究中深入思考。在技术关联网络的应用方面,可考虑加入时间元素构建时序技术关联网络,描述不同时期技术关联情况,揭示技术关联的产生、发展和消亡的过程,分析技术前沿、热点与发展趋势;了解技术研究现状、发展脉络和最新态势,科学制定发展战略规划,规避科研投资风险并正确选择技术路径;等等。

6 结语

本文提出了一种基于问题－技术矩阵的科技文献中技术关联发现方法。通过提取科技文献中的研究问题和技术形成初始问题－技术矩阵,为降低矩阵数据稀疏性对技术条目进行聚类,然后分析发现技术间存在直接竞争关系、间接竞争关系和合作关系。通过实验验证技术关联发现方法的有效性,实验以 AAAI 会议集为研究对象,从科技文献中研究问题－方法矩阵出发对技术间关系进行了归纳和总结,探索自动化建立技术之间的关联关系的方法,为技术关联自动发现提供一种新思路。与之前从专利、企业角度对技术关联进行研究的工作相比,本文从科技文献中的研究问题角度对技术间关系进行了归纳和总结,自动化发现技术间关系发现基于问题－技术矩阵的技术关联关系分析在一定程度上可以辅助科技创新管理与工作。本项研究是面向科技文献的技术关联发现与应用的探索性研究,研究结果有助于提高科技创新研究与管理的工作效率。

受时间和能力所限,本文研究还存在一些不足,未来工作展望如下:①技术、问题的抽取自动化水平有待提高,技术抽取受限于技术词表的构建,覆盖率和新颖性有待改进,问题模板库的更新维护等问题需要结合语料库不断更新维护,未来工作可考虑将新技术术语发现和技术词表结合的方法满足技术监测新颖性需求;②当前研究中仅总结了四种技术关联关系发现方法,然而实际应用中技术关联种类更多,未来工作可尝试分析发现更多技术关联关系,并通过推理推断提高自动化构建水平;③技术关联网络构建是在 AAAI 会议论文集进行实验,所得出的结论受限于数据集的范围,尚未在医药、化学、传统石油等更多行业进行实证,后续工作可考虑扩大数据集规模及领域范围,验证扩展本文所提方法的通用性。

参考文献:

[1] BHARADWAJ A , SAWY O A E , PAVLOU P A , et al. Digital business strategy: toward a next generation of insights[J]. Mis quarterly, 2013, 37(2):471－482.
[2] 国家发展改革委. 战略性新兴产业形势判断及"十四五"发展建议[EB/OL]. [2021－05－04]. https://baijiahao.baidu.com/s? id=1688663466446852041&wfr=spider&for=pc.
[3] 吴江宁, 张红卫, 王舒. 基于科技文献的时序主题链构建方法[J]. 科学学研究, 2014, 32(9):1306－1312.
[4] 王佳琪,张均胜,乔晓东. 基于文献的科研事件表示与语义链

接研究[J]. 数据分析与知识发现,2018,2(5):32－39.
[5] 张昱,张均胜,姚长青. 科技文献语义关联网络及其应用探析[J]. 中国科技资源导刊,2020,52(1):40－47.
[6] 王艳艳,张均胜,乔晓东,等. 基于问题——方法矩阵的文献新颖性评估方法[J]. 情报理论与实践,2021,44(2):90－95.
[7] LI J, TANG J, ZHANG J, et al. Arnetminer: expertise oriented search using social networks[J]. Frontiers of computer science in China, 2008, 2(1):94－105.
[8] 张兆锋,张均胜,姚长青. 一种基于知识图谱的技术功效图自动构建方法[J]. 情报理论与实践,2018,41(3):149－155.
[9] 陈颖,张晓林. 专利技术功效矩阵构建研究进展[J]. 现代图书情报技术,2011(11):1－8.
[10] 王丽,张冬荣,张晓辉,等. 利用主题自动标引生成技术功效矩阵[J]. 现代图书情报技术,2013(5):80－86.
[11] 陈颖,张晓林. 基于特征度和词汇模型的专利技术功效矩阵结构生成研究[J]. 现代图书情报技术,2012(2):53－59.
[12] 张博培. 面向专利的术语识别与技术功效矩阵构建技术[D]. 北京:北京工业大学,2015.
[13] 翟东升,张京先,胡等金. 基于 SAO 结构和词向量的专利技术功效图自动构建研究[J]. 情报理论与实践,2020,43(3):116－123.
[14] 许海云,方曙. 基于专利功效矩阵的技术主题关联分析及核心专利挖掘[J]. 情报学报,2014,33(2):158－166.
[15] 刘化然,曹旭,张晓冬,等. 基于专利技术功效矩阵的技术机会识别方法[J]. 图书情报导刊,2020,5(6):65－70.
[16] 栾春娟,刘则渊,王贤文. 发散与收敛:技术关联度的演变趋势分析——以全球太阳能技术的专利计量为例[J]. 研究与发展管理,2013,25(4):87－95.
[17] 丰雷. 简论产业集群与技术关联关系[J]. 管理与财富,2009(12):64－65.
[18] 姜红. 基于技术关联性视角的产业创新模式与技术选择理论研究[D]. 长春:吉林大学,2008.
[19] HOFMANN P, KELLER R, URBACH N. Inter-technology relationship networks: arranging technologies through text mining[J]. Technological forecasting and social change, 2019, 143(June): 202－213.
[20] 冯秀珍,李萌,刘俊婉. 面向技术用途的关联关系识别方法研究[J]. 科技进步与对策,2014(9):19－23.
[21] 丰雷. 产业集群与技术关联关系研究[J]. 科技和产业,2010,10(5):4－6,21.
[22] 陈卫静,刘静,凌世婷. 媒介角色理论视域下技术间关联关系研究——以电子信息领域的高影响力论文为例[J]. 图书情报工作,2017,61(22):107－116.
[23] 黄斌,黄鲁成,吴菲菲,等. 基于专利共类的技术间关联性评估[J]. 情报杂志,2015(2):99－103.
[24] 黄斌. 技术关联关系研究[D]. 北京:北京工业大学,2016.
[25] 杨冬敏. 基于关联关系的技术演化研究[D]. 武汉:武汉大学,2019.
[26] 王燕鹏. 基于动态主题模型的机构研究主题分布及演化[D]. 北京:中国科学院文献情报中心,2017.
[27] 李勇敢. 技术领域维度下相对技术关联度研究——以德温特专利数据库共类分析为例[J]. 科技进步与对策,2017,34(7):146－153.
[28] 黄晓斌 梁辰. 专利技术的关联网络分析——以 4G 通信技术领域为例[J]. 情报学报, 2015,34(1): 92－104.
[29] 包翔,刘桂锋. 一种基于概率潜在语义分析的专利主题标引方法研究[J]. 情报工程,2020,6(3):15－24.
[30] 汪锦霞,刘向. 基于 LOGISTIC 拟合的动态技术轨道识别与评价研究[J]. 情报工程,2020,6(3):63－77.
[31] 杨金庆,陆伟,吴乐艳. 面向学科新兴主题探测的多源科技文献时滞计算及启示——以农业学科领域为例[J]. 情报学报,2021,40(1):21－29.

作者贡献说明:
徐珍珍:论文起草;采集、清洗数据,实验分析;论文修订;
张均胜:提出研究思路,设计,论文修订、审核;
刘文斌:数据清洗,实验实现。

Automatically Discovering Associations Among Technologies in Scientific Literature

Xu Zhenzhen Zhang Junsheng Liu Wenbin
Institute of Scientific and Technical Information of China, Beijing 100038

Abstract: [**Purpose/significance**] Technology association analysis can help government to make science and technology strategies and plans, and help enterprises to make up shortcomings for development, and help researchers to select research directions. [**Method/process**] This paper proposed an approach to discover technology associations and relations by constructing a problem-technology matrix based on scientific literature. First, research questions and technologies were extracted from contents of scientific literature to formulate a question-technology matrix; and then, technologies were calculated to cluster for reducing data sparsity. Finally, experiments showed the effectiveness of our proposed approach. [**Result/conclusion**] Starting from the question-technology matrix in literature, this paper summarizes the associations between technologies, and explores methods of automatically establishing the relationship between technologies, which can effectively help to discover the relationship between technologies, such as direct competition relationship, cooperation relationship, indirect competition relationship and cooperation relationship. This research is an exploratory research on the discovery and application of technology association for scientific literature. The research results will help to improve the efficiency of scientific and technological innovation research and management.

Keywords: question-technology matrix technology relationship technology similarity